August 27, 2024

Honorable Ona T. Wang
United States Magistrate Judge
Southern District of New York

Re: *The New York Times Company v. Microsoft Corporation, et al.*,
Case No.: 23-cv-11195: Dispute Regarding OpenAI's Search Terms

Dear Magistrate Judge Wang:

Plaintiff The New York Times Company ("The Times") requests a conference to discuss its request for OpenAI to apply additional search terms to its custodians' ESI.[1] Exhibit 1 compares The Times's proposed terms with the terms that OpenAI initially offered and the document hit counts for each set.[2] The Times's "search terms are reasonable, the burden is not excessive, and there is no unfairness." *Giuffre v. Dershowitz*, 2022 WL 827825, at *3 (S.D.N.Y. Feb. 14, 2022).

1. OpenAI Recycled Terms from Other Cases and Refuses to Apply Relevant Terms.

The agreed-upon provisions of the ESI Order contemplate cooperation on search terms. Dkt. 135-2 ¶ 5.[3] Since June 14, The Times has made *three* search term proposals to OpenAI, each time narrowing its proposal based on hit reports provided by OpenAI. *See* Ex. 2; Dkt. 204-1. Yet ***OpenAI has rejected every single search term*** proposed by The Times, including terms that hit on as few as *three* documents. OpenAI, in effect, is refusing to consider any input from The Times about appropriate search terms to locate relevant documents. Indeed, with one exception, each term will capture documents that OpenAI has already agreed to produce.[4]

Worse, the terms that OpenAI proposed were clearly designed for other cases, including the *Authors Guild* class action, which asserts claims on behalf of book authors. *See, e.g.*, Ex. 1, Section H (proposing book-related terms including "z-library" and "books corpus"). As with its insufficient offer of custodians (Dkts. 204, 211), OpenAI's approach to search terms contradicts its assertion that this case is "very different" from *Authors Guild*. Dkt. 72 at 16.

The only concrete objection OpenAI has raised is that The Times's proposals purportedly capture too many documents. Ex. 2. That objection is unfounded. The proposal is reasonably tailored to capture documents about issues that, in Microsoft's words, make this case "substantially broader" than other cases against OpenAI, Dkt. 153 at 2, yielding approximately 546,902 additional documents[5] (as compared with OpenAI's proposal, which captures 131,601

[1] The parties conferred by videoconference on July 31, and through numerous written exchanges. Ex. 2; Dkt. 204-1. OpenAI has exhibited a pattern of refusing to make any meaningful concessions during party negotiations in this case, instead waiting for The Times to file a motion to compel and then complaining it is "premature." *See* Dkt. 211 at 1. This motion may be no different, but given OpenAI's refusal to negotiate—thereby consuming two months of a six-month discovery period—The Times has little option but to again move to compel.

[2] Terms will be referred to by the numbers in Ex. 1, Column A—*e.g.,* Term 1 is "paywall," Term 2 is "antiregurg*."

[3] The parties have a dispute about whether the ESI order should include a cap on the number of document custodians per side. The Times opposes any artificial cap. Dkt. 135. The parties otherwise agree on all aspects of the ESI order.

[4] The sole exception is for Terms 59-60, which relate to licensing agreements that are the subject of The Times's motion to compel OpenAI to produce custodial documents. Dkt. 141.

[5] While Exhibit 1 lists 546,902 as the number of documents captured by The Times's proposal, the number of documents is actually slightly lower because The Times dropped one term from this motion.

documents).[6] Moreover, OpenAI has never explained why it cannot run at least some of the terms proposed by The Times, including those that hit on a very small number of documents.[7]

2. The Times Has Proposed Search Terms Related to Issues Central to This Case.

a. Terms Related to the Outputs of Defendants' Products

The Times proposes several terms designed to capture documents related to the outputs of Defendants' products—*e.g.*, how ChatGPT responds to user queries. Ex. 1, Section A. Despite arguing elsewhere that these outputs are central to this case, *see* Dkt. 72 ¶¶ 41-42, OpenAI has run only four search terms on this issue, which have yielded paltry document hit counts (approximately 3,000 combined hits). The Times's proposal includes terms related to the technology underlying generative search known as retrieval augmented generation ("RAG"), which enables Defendants' products to display extensive excerpts or paraphrases of Times content.[8] *See, e.g.*, FAC ¶¶ 81, 108-23, 163. *See* Dkt. 153 at 2-3 (discovery on "generative search technology" will be "significant").

Relatedly, The Times's proposal includes "paywall" and terms focused on "regurgitation," which relate to how Defendants' products enable users to circumvent paywalls and generate verbatim copies of copyrighted content. FAC ¶¶ 102–23, 157.[9] These terms are likely to yield additional responsive documents, as evidenced by OpenAI's productions to date. *See* Ex. 3 at -03 (email discussing a [redacted]).

b. Terms Related to GenAI and Journalism

The Times proposes several terms targeted at the relationship between GenAI and journalism, including harm to The Times. Ex. 1, Section B; *see* FAC ¶¶ 47-48, 154-57 ("GenAI products threaten high-quality journalism"). OpenAI should run "NY Times" (and other iterations), particularly given the low hit counts. Ex. 1, Term 11. The "NY Times" terms will capture, among other documents, information related to the [redacted]. Ex. 4 at -27. These terms also capture documents about the diversion of traffic from The Times and other publishers. *E.g.*, Term 13. OpenAI has run only two search terms related to these issues (11 and 33), which again hit on a paltry number of documents (roughly 5,000).

c. Terms Related to The Times's Legal Claims

The Times proposes terms designed to capture documents related to copyright infringement, trademark infringement, and the DMCA claims. Ex. 1, Section C; *see* FAC ¶¶ 158-91. These terms will test Defendants' assertions that Defendants genuinely believed their conduct constituted fair use. Dkts. 52 at 7; 65 at 16. *See also* Ex. 5 at -71 [redacted]

---

[6] The Times has also moved to compel OpenAI to produce documents from more custodians. Dkt. 204. That dispute has no bearing on this one, including because, in violation of the agreed portion of the ESI Order, OpenAI refuses to provide hit counts for those additional custodians. Dkt. 204-1 at 10; Dkt. 135-2 ¶ 5. OpenAI has only provided updated hit counts for the four additional custodians it agreed to add on July 31. *See* Dkt. 211; Ex. 2. Moreover, those updated hit counts were only run against The Times's proposed terms—not the search terms OpenAI initially agreed to run (as reflected in Exhibit 1, Column E). As such, the hit counts presented in Exhibit 1 is the only set of hit counts that compares the parties' competing terms to the same set of custodians—specifically, the 12 custodians initially selected by OpenAI.

[7] During a July 31 call, OpenAI for the first time asked The Times to provide a list that specifies, for *each* proposed search term, the particular RFP that applies to that term. The Times rejected this request because, among other reasons, OpenAI never once during the prior six weeks raised any relevance objection to any proposed term.

[8] OpenAI initially refused to produce documents related to RAG but changed its position after The Times filed a motion to compel on this issue. Dkts. 141, 147.

[9] *See also OpenAI and Journalism*, OPENAI (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/ (describing problems with memorization and regurgitation).

[redacted] The terms "copy* /5 shield" (a reference to OpenAI's Copyright Shield policy) and "indemnify," address OpenAI's offer "to indemnify ChatGPT customers for copyright infringement,"[10] which is relevant to OpenAI's argument that "normal people do not use OpenAI's products" to elicit copyrighted content. Dkt. 52 at 2. Finally, Terms 21-24 will capture documents about how OpenAI could have prevented the products from infringing copyrights.

d. Terms Related to Model Training

The Times proposes terms related to the training of the models for Defendants' products. Ex. 1, Section D. OpenAI agrees this topic is relevant and has provided search terms for it, but most of the hits stem from one term designed for *Authors Guild*. *See* Term 25 (referencing "book"). The Times's proposed terms are appropriately broader, tracking all phases of the training process, including pre-training, post-training, and fine tuning. *See* Dkt. 136-1 at 1 (OpenAI agreeing to produce documents "to cover the entire training process"). Relatedly, The Times's terms capture documents relating to how Defendants prioritize high-value content for training. *E.g.*, Terms 25-27; FAC ¶ 87 (alleging that Defendants prioritized Times content); *see also* Ex. 6 at -07 (OpenAI employees describing [redacted]). These terms also include datasets that Defendants mined for content to train their models, including "webtext," "common crawl," and C4, and The Times's proposal accounts for various iterations of these terms (e.g., "WT" for "webtext", and the "CC" term for common crawl). FAC ¶¶ 87-91; Terms 28-32. Finally, The Times's terms also include codenames for OpenAI's models, without which Defendants' productions are likely to be incomplete. *E.g.,* Terms 28, 33 (including "dv3" and "davinci-3", [redacted]); Ex. 7 at -01.

e. Terms Related to Technical Documentation and Processes

The Times proposes terms designed to capture documents relating to technical specifications of the models and products. Ex. 1, Section E. "Azure" refers to the supercomputing systems powered by Microsoft's cloud computer platform, which were used to train all of OpenAI's GPT models after GPT-1. FAC ¶ 68. The proposal also includes various codenames that have been revealed through OpenAI's productions so far, including [redacted]. Ex. 8 at -14.

f. Terms Related to Collaboration with Microsoft

The Times proposes terms designed to capture documents related to OpenAI's collaboration with Microsoft. Ex. 1, Section F. *See* Dkt. 147 at 3 (OpenAI claiming that it is "producing communications with Microsoft" in response to RFP 15). The proposal includes codenames revealed through document productions [redacted]. *See* Ex. 9 at -74; Ex. 10 at -10.

g. Terms Related to the Market for and Value of Content

The Times proposes terms related to the market for and value of Times content. Ex. 1, Section G. Term 61 is designed to capture documents about how OpenAI could have used non-copyrighted material to train its models. Dkt. 147 at 2 (agreeing to produce documents regarding this issue). Terms 59-60 are designed to capture documents concerning licensing agreements.

---

[10] Slaughter and May, *OpenAI Offers to Indemnify ChatGPT Customers for Copyright Infringement*, THE LENS (Nov. 14, 2023), https://www.lexology.com/library/detail.aspx?g=671fdd7f-3cef-4606-bb40-6f1c3dbaefe0.

Respectfully submitted,

/s/ Ian B. Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

/s/ Steven Lieberman
Steven Lieberman
Rothwell, Figg, Ernst & Manbeck

cc: All Counsel of Record (via ECF)